Submitted to Neural Information Processing Systems (NIPS) 1999.

Multi-Parent Clustering Algorithms from Stochastic Grammar Data Models

Eric Mjolsness
Jet Propulsion Laboratory
California Institute of Technology
Pasadena CA 91109-8099
mjolsness@jpl.nasa.gov

Rebecca Castaño
Jet Propulsion Laboratory
California Institute of
Pasadena CA 91109-8099
becky@aig.jpl.nasa.gov

Alexander Gray
Jet Propulsion Laboratory
California Institute of Technology
Pasadena CA 91109-8099
agray@aig.jpl.nasa.gov

Abstract

We introduce a statistical data model and an associated optimization-based clustering algorthm which allows data vectors to belong to zero, one or several "parent" clusters. For each data vector the algorithm makes a discrete decision among these alternatives. Thus, a recursive version of this algorithm would place data clusters in a Directed Acyclic Graph rather than a tree. We test the algorithm with synthetic data generated according to the statistical data model. We also illustrate the algorithm using real data from large-scale gene expression assays.

1 Introduction

Clustering algorithms traditionally construct clusters which are related by placement in a tree (hierarchical clustering) or embedding in a low-dimensional space (self-organizing maps). We seek to generalize the deterministic annealing approach to clustering under mixture models [1][2][3] so that when used recursively it can construct a Directed Acyclic Graph (DAG) of clusters rather than a tree. As a key development to this end, we consider here the recursively applicable step of clustering a set of feature vectors into groups so that each vector belongs to zero, one, or two clusters. Thus each data vector can have multiple "parent" clusters. We report on a generative model for such data using stochastic parameterized grammars, derive an appropriate constrained optimization problem for inferring parent clusters from data, and define and test a suitable optimization algorithm for this problem.

2 Theory

2.1 DataModel

A generative, statistical data model is defined using stochastic parameterized grammars [4] as follows. A stochastic grammar has a "start" symbol, and a unique rule which transforms this symbol into a level-zero "cluster" symbol with numerical parameters including a mean and (possibly diagonal or scalar) covariance which are specified by the grammar rather than generated by a probability distribution. This level-zero cluster can serve as the left hand side for several different rules. One rule simply destroys the cluster. Another permits it to create data vectors according to its mean and covariance; these model "distractor" data which do not belong to any actual (level-one) cluster. Finally, a level-zero cluster may generate a level-one cluster whose level-one mean is determined by drawing from the level-zero Gaussian distribution, and whose level-one covariance is specified by the grammar. (Alternatively one may use a prior such as a cut-off inverse power law for diagonal covariance entries.) The level-zero cluster survives this rule-firing event and can participate in further rule firings. For spherical Gaussians, this rule may be summarized as:

$$\text{Cluster0}(y,\sigma,c) \rightarrow \text{Cluster0}(y,\sigma,c+1),\ \text{Cluster1}(y',\sigma',c,k)$$

$$E = \frac{1}{2\sigma_0^2}\|y - y'\|^2 - \mu_0$$

Here y and y' are mean feature vectors, E is an energy function whose Boltzmann probability distribution contributes to the stochastic behavior of the grammar as described below. Based on the relative probabilities of these three rules in the stochastic grammar, the level-zero cluster generates some number of level-one parameterized clusters and distractors, and then dies.

The level-one clusters also can serve as the left-hand-side of several rules in the stochastic parameterized grammar. One rule kills the cluster. One rule allows it to generate a data vector (interpreted as a real cluster member, not a distractor) according to a Gaussian using the cluster's mean and covariance; the cluster symbol survives as well. And one special rule takes *two* clusters on its left-hand-side and generates a single data vector by a suitably weighted average of the parent mean and covariance parameters. Both parents survive the event. This rule is the origin of multiple parentage in the data model. For scalar covariance (spherical Gaussians), this rule may be summarized as:

$$\text{Cluster1}(y,\sigma,c,k),\ \text{Cluster1}(y',\sigma',c',k') \rightarrow$$

$$\text{Cluster1}(y,\sigma,c,k+1),\ \text{Cluster1}(y',\sigma',c',k'+1),$$

$$\text{Datum}(\text{x},(c,k+1),(c',k'+1))$$

$$E = \frac{1}{2\sigma_1^2}\|x_i - ((y+y')/2)\|^2 - \mu_1$$

Finally, as discussed for a previous single-parent stochastic grammar [1], a global permutation removes all identifying indices (c,k) from all the generated data vectors.

Each rule in the grammar has an energy function which induces an unnormalized Boltzmann probability factor $\exp(-E/T)$. By analogy with statistical mechanics, we take the probability of an entire derivation ∂ through the grammar to be the product of the Boltzmann factors for the all rules that fired in the derivation, normalized by the partition function which is the sum of all such products:

$$\Pr = \exp(-\frac{1}{T}\sum_{r\in\partial} E_r)/Z$$

$$Z = \sum_{\partial}\exp(-\frac{1}{T}\sum_{r\in\partial} E_r)$$

From this distribution one can derive Bayesian inference algorithms for the cluster means and covariances given data generated by the stochastic parameterized grammar.

2.2 Objective function

If we let i index the observed data vectors and α index the clusters, then we can record which data vectors were generated by which parents using two arrays N with different numbers of indices:

$N_{i\alpha}$ = 1 if i was generated by α alone, zero otherwise

$N_{i\alpha\beta}$= 1 if i was generated by α and β, zero otherwise.

The simplest objective function for inferring N and the cluster means y from data x which we can derive from the grammar as outlined above using the methods of [4][5] is:

$$E = \sum_{i\alpha} N_{i\alpha}[\frac{1}{2\sigma_1^2}\left\|x_i - y_\alpha\right\|^2 - \mu] + \sum_{i\alpha\beta} N_{i\alpha\beta}[\frac{1}{2\sigma_1^2}\left\|x_i - \left((y_\alpha + y_\beta)/2\right)\right\|^2 - \nu]$$

with constraints

$$N_{i\alpha} \in \{0,1\}, N_{i\alpha\beta} \in \{0,1\},$$
$$\sum_{\alpha} N_{i\alpha} + \sum_{\alpha\beta} N_{i\alpha\beta} \le 1$$

We can reduce the number of variables to solve for by noticing that a change of variables is possible:

Define $M_{i\alpha}$ = 1 if i was generated by α alone or in concert with another cluster, zero otherwise. Then

$$N_{i\alpha} = M_{i\alpha}(2 - S_i)$$
$$N_{i\alpha\beta} = M_{i\alpha}M_{i\beta} - \delta_{\alpha\beta}M_{i\alpha}\,.$$
$$S_i = \sum_{\alpha} M_{i\alpha}$$

The objective function becomes

$$E = \sum_{i\alpha} M_{i\alpha}[\frac{1}{2\sigma_1^2}\left\|x_i - y_\alpha\right\|^2 - \mu] - (\nu - \mu)\sum_i S_i(1 - S_i)$$
$$- \sum_{i\alpha\beta} M_{i\alpha}M_{i\beta}\frac{1}{2\sigma_1^2}\left\|(y_\alpha - y_\beta)/2\right\|^2 \qquad (1)$$

with constraints

$$M_{\alpha i} \in \{0,1\},$$
$$\sum_{\alpha} M_{\alpha i} \leq 2$$

A similar change of variable applies for data generated by up to three parents:

$$\begin{aligned}
N_{i\alpha} &= \frac{1}{2} M_{i\alpha}(S_i - 2)(S_i - 3) \\
N_{i\alpha\beta} &= (M_{i\alpha} M_{i\beta} - \delta_{\alpha\beta} M_{i\alpha})(3 - S_i) \\
N_{i\alpha\beta\gamma} &= M_{i\alpha} M_{i\beta} M_{i\gamma} - \delta_{\alpha\beta} M_{i\alpha} M_{i\gamma} - \delta_{\beta\gamma} M_{i\alpha} M_{i\beta} \\
&\quad - \delta_{\alpha\gamma} M_{i\beta} M_{i\gamma} + 2\delta_{\alpha\beta}\delta_{\alpha\gamma} M_{i\alpha} \\
S_i &= \sum_{\alpha} M_{i\alpha}
\end{aligned}$$

We turn now to the construction of iterative algorithms for optimizing under these objective functions and constraints.

2.3 Algorithm

To perform the constrained optimization, we may consider multi-parent clustering as a modification of the existing soft-max style clustering in which the WTA (winner-take-all) or WMTA (winner might take all) constraint is replaced with n-winners by means of a dual encoding of a membership M and its complement $\overline{M} = 1 - M$. If α indexes the clusters and i indexes the data, then

$$\sum_{\alpha} M_{\alpha i} \leq n,$$

which can be implemented via

$$\begin{aligned}
&M_{\alpha i} + \overline{M}_{\alpha i} = 1 \\
&M_{\alpha i}, \overline{M}_{\alpha i}, s \geq 0. \\
&\sum_{\alpha} M_{\alpha i} + s = n
\end{aligned}$$

The latter three lines can be translated into alternative soft-max objective functions analogous to a Mean Field Theory Potts glass effective energy:

$$\begin{aligned}
E = &\sum_{\alpha i} M_{\alpha i}\left(D_{\alpha i} - \mu\right) + T\sum_{\alpha i} M_{\alpha i}\left(\log M_{\alpha i} - 1\right) + T\sum_{\alpha i} \overline{M}_{\alpha i}\left(\log \overline{M}_{\alpha i} - 1\right) \\
&+ T\sum_{i} s_i\left(\log s_i - 1\right) + \sum_{\alpha i} \nu_{\alpha i}\left(M_{\alpha i} + \overline{M}_{\alpha i} - 1\right) + \sum_{i} \lambda_i\left(\sum_{\alpha} M_{\alpha i} + s_i - n\right)
\end{aligned}$$

Here D includes the distance metric in the first term of the objective function (1), and can also locally reflect the quadratic terms in (1) in an iterative algorithm as is done in the soft-assign approach to quadratic assignment optimization [3][7]. Taking derivatives of E with respect to each type of variable, and initializing the

Lagrange multipliers to zero, we can derive aggressive update dynamics (large descent steps) similar to the soft-assign algorithm [6][3]:

$$M^{0}_{\alpha i} = \exp\left[\left(\mu - D_{\alpha i}\right)/T\right];$$
$$\overline{M}^{0}_{\alpha i} = 1;$$
$$s^{0}_{i} = 1;$$

and then iteratively,

<

$$M'_{\alpha i} = kM^{prev}_{\alpha i} / (s^{prev}_{i} + \sum_{\beta} M^{prev}_{\beta i});$$
$$s_{i} = ks^{prev}_{i} / (s^{prev}_{i} + \sum_{\beta} M^{prev}_{\beta i});$$
$$M_{\alpha i} = M'_{\alpha i} / (M'_{\alpha i} + \overline{M}^{prev}_{\alpha i});$$
$$\overline{M}_{\alpha i} = \overline{M}^{prev}_{\alpha i} / (M'_{\alpha i} + \overline{M}^{prev}_{\alpha i});$$

>

Then the cluster means are updated and the above steps iterated, with occasional decreases in temperature according to a fixed e.g. exponential schedule. If this algorithm failed to converge it would be possible to back off and do the gradient descent steps more slowly than the constraint-enforcement (ascent-like) steps. However our numerical experiments show good convergence in the present context. Convergence theory for soft-assign is dealt with in [7][8].

The number of clusters could be varied with repeated runs, e.g. so as to produce a reasonably small average data-to-cluster-center distance without too many cluster centers and a relatively high likelihood for the overall clustering as measured by cross-validation [9].

3 Results

3.1 Data

To demonstrate the algorithm, we show an example using two clusters of data points generated from 2-D Gaussian distributions. One third of the points were generated from a zero mean Gaussian. Another third were generated from a Gaussian centered at (10,10). The final third were generated from a combination of the first two Gaussians.

3.2 Clustering

The results shown are the estimated means and the probability of several types of errors. There are three types of errors that can occur.:

(1) A point that should be classified as coming from both clusters could be assigned to one or no clusters.

(2) A point that comes from one cluster may be assigned to no clusters.

(3) A point that comes from one cluster could be assigned to two clusters.

The probabilities for each of these types of errors is plotted in Figure 1as a function of the parameter μ, which represents a reward for being assigned to a cluster. In these experiments we have taken $V = \mu$. From the figure, it can be seen that when the reward for joinning a cluster is too small, the points that should belong to only one cluster tend to be assigned to no"clusters. As the reward for joining a cluster gets larger, all points are assigned to two clusters.

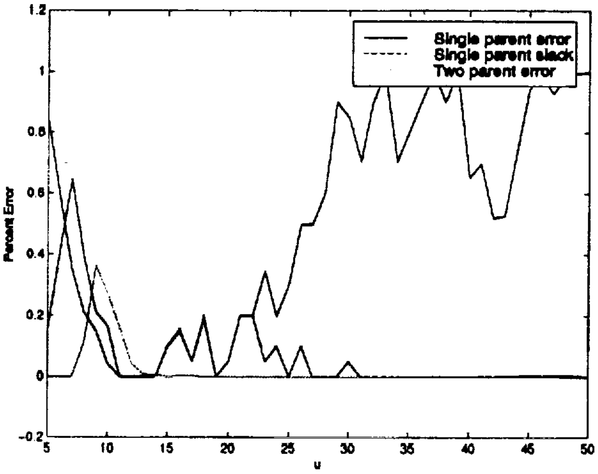

Figure 1. Three types of error as a function of μ.

In Figure 2, we show the estimated means for the two clusters as a function of the parameter μ. The average value of the first coordinate for each cluster is plotted. For μ too large or too small the means tends toward each other and the joint mean of the entire data set. There is an intermediate window of successful operation.

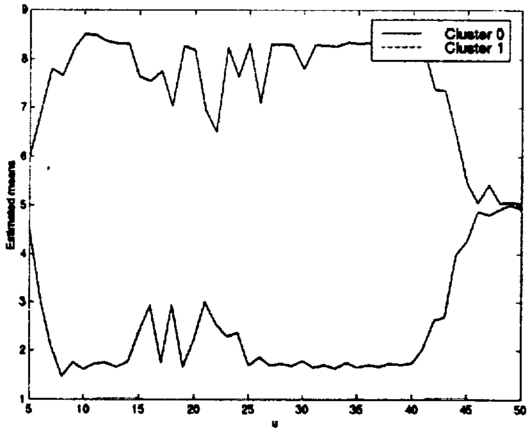

Figure 2. Cluter means (first coordinate) as a function of μ.

We are currently performing a similar parameter exploration for higher–dimensional feature vectors and more classes. So far we observe a reasonable rate of successful multiparent clustering runs for 15 clusters in 10 dimensions.

In addition, we have also run the algorithm on real biological data consisting of 1244 feature vectors, each truncated to 5 dimensions, representing log ratios of mRNA gene expression measurements from Stuart Kim's laboratory on the nematode worm *C. elegans*. We used 15 clusters, and varied μ. Depending on the value of μ, we observe varying fractions of genes falling into the slack class, having single parent clusters, and having two parent clusters.

4 Discussion

We have introduced a statistical data model and an optimization algorithm for analysing clustered data in which a data vector can belong to zero, one, or several clusters. This "multiparent clustering" algorithm, applied recursively, would create a Directed Acyclic Graph rather than a tree of hierarchical cluster centers. For many applications in data analysis, visualization and information retrieval the DAG is a more reasonable or flexible structure to infer. We demonstrated the algorithm using synthetic data generated according to the multi-parent clustering statistical data model.

Acknowledgments

This work was supported in part by the NASA Autonomy program, the Whittier Foundation, and the Office of Naval Research under contract N00014-97-1-0422. Stuart Kim's Stanford University laboratory provided the *C. elegans* gene expression array data.

References

[1] R. A. Redner and H. F. Walker, "Mixture Densities, Maximum Likelihood, and the EM Algorithm", SIAM Review 26:195-239 1984.

[2] R. J. Hathaway, "Another Interpretation of the EM Algorithm for Mixture Distributions", Statisticas and Probability Letters 4:53-56, 1986.

[3] Steven Gold, Anand Rangarajan, and Eric Mjolsness, "Learning with Preknowledge: Clustering with Point and Graph Matching Distance Measures", Neural Computation, vol 8 no 4, May 15 1996.

[4] Eric Mjolsness, "Connectionist Grammars for High-Level Vision", in Artificial Intelligence and Neural Networks: Steps Toward Principled Integration, eds. Vasant Honavar and Leonard Uhr, Academic Press, 1994.

[5] Eric Mjolsness, "Symbolic Neural Networks Derived from Stochastic Grammar Domain Models", in *Connectionist Symbolic Integration*, eds. R. Sun and F. Alexandre, Lawrence Erlbaum Associates, 1997.

[6] Anand Rangarajan, Steven Gold, and Eric Mjolsness, "A Novel Optimizing Network Architecture with Applications" Neural Computation, vol 8 no 5, July 1996.

[7] Anand Rangarajan, Alan Yuille, Steven Gold and Eric Mjolsness, "A Convergence Proof for the Softassign Quadratic Assignment Algorithm", *Advances in Neural Information Processing Systems 9*. M. Mozer, M. Jordan, and T. Petsche, eds. MIT Press, 1997.

[8] A. Rangarajan, A. Yuille, and E. Mjolsness. Neural Computation, Convergence Properties of the Softassign Quadratic Assignment Algorithm", to appear 1999.

[9] P. Smyth, "Clustering using Monte Carlo Cross-Validation", Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1996.